Special Collection

Active Learning Training Strategy for Predicting O Adsorption Free Energy on Perovskite Catalysts using Inexpensive Catalyst Features

Shambhawi Shambhawi,[a] Gábor Csányi,[b] and Alexei A. Lapkin*[a, c]

Machine learning (ML) based energy prediction models are among the most effective descriptor-based catalyst screening tools for heterogeneous reaction systems. However, their implementations are limited due to expensive data labelling, ab initio feature evaluation and lack of universal catalyst features, that is, beyond d-band theory. Herein, we propose an inexpensive geometric feature for application on systems beyond d-band theory, for example perovskites comprising of s-, p-, d- and f-block elements. We outline a workflow that inputs these features into an active learning algorithm that enables effective data labelling, whilst improving prediction accuracies of existing models. We then use batch sampling to define termination criteria and to implement time-series error forecasting for further reducing the number of expensive data labelling for training. We implement this workflow to train ML models for predicting oxygen adsorption free energy on perovskites and achieve similar, if not better, prediction accuracies as obtained from *ab initio* features.

1. Introduction

A catalyst's performance for a given reaction system can be predicted using microkinetic models that are based on ab initio methods like density functional theory (DFT).[1] However, their implementation as a catalyst screening tool is limited due to the expensive quantum chemical calculations. A common technique to circumvent this limitation is to identify reaction descriptors that demonstrate high correlation with the overall product formation rate, for example binding energies of a relevant intermediate.[2] This descriptor is then evaluated on different catalyst surfaces to further compare their performance, thus reducing the computationally expensive model evaluation to a cheap descriptor evaluation.[3] Nonetheless, DFT evaluation of a descriptor for more than a thousand of catalysts, including different facets and adsorption sites, still requires significant computational time and resources. Therefore, developing a high-throughput screening tool for prioritizing experimental and theoretical studies is among the most fundamental goals in heterogeneous catalysis today.

The idea of a screening tool is to provide a near estimate value of the reaction descriptor at the least computational expense possible. A number of approaches has been reported in the literature for this purpose, starting with the linear scaling relations between the descriptor molecule and the properties of the binding atom,[4] to regression models for screening catalysts from a limited search space, like bi-metallics.[5] Finally, the graph based convolutional neural nets (GCNN) for a complete generalized screening have been introduced.[6]

Despite promising applications, all these methods have their limitations. Scaling relations are based on the d-band theory[7] and can perform terribly on catalysts comprising of s-, p-, d- and f-bock elements, like perovskites. The regression models, on the other hand, are developed for limited search spaces that have similar adsorption sites. They also employ expensive *ab initio* catalyst features, like band-width, for training models that makes the entire process of developing a model redundant.[5b,8] Although, some studies[5a,c] have reported prediction models for intermetallic systems (active d-band) that use features derived from the inexpensive semi-empirical relations of the linear muffin-tin orbital theory (LMTO).[9] Lastly, the GCNN approach is limited due to its extensive training data requirement. The data requirement could vary from a few hundred (for single adsorption sites) to tens of thousands (generalized prediction).[6a] Even though large-scale datasets generation projects have been reported, the recent one being Open Catalyst 2020,[10] the dataset contains binding energies of only 0.07 % of the total feasible catalytic surfaces calculations.[10]

[a] *S. Shambhawi, Prof. A. A. Lapkin*
Department of Chemical Engineering and Biotechnology
University of Cambridge
Cambridge CB3 0AS (UK)
E-mail: aal35@cam.ac.uk
[b] *Prof. G. Csányi*
Department of Engineering,
University of Cambridge
Trumpington Street
Cambridge CB2 1PZ (UK)
[c] *Prof. A. A. Lapkin*
Cambridge Centre for Advanced Research and Education in Singapore Ltd
1 Create Way
CREATE Tower #05-05
138602 (Singapore) Singapore

Supporting information for this article is available on the WWW under https://doi.org/10.1002/cmtd.202100035

This publication is part of a joint Special Collection of Chemistry-Methods and the European Journal of Organic Chemistry including contributions focusing on "Automating Synthesis: From Planning to Execution". Please visit chemistry-methods.org/collections to view all contributions.

Despite their limitations, machine-learning (ML)-based prediction models have become popular as a data processing method for high-throughput materials screening.[11] Given a defined search space, regression models are usually the common choice for developing screening tools, due to their ease of implementation and a relatively small data requirement as compared to GCNN.[6a] Studies[5a,c] also report better prediction accuracies of these models, compared to other methods, whilst employing LMTO based features for intermetallic systems. Other studies[8b,12] focus on further reducing the required training data by implementing a technique called 'active learning'. This technique is very useful in situations where unlabeled data is abundant and data labeling is expensive, like obtaining binding energies of intermediates on thousands of catalysts surfaces using DFT. Within this technique a learning algorithm actively queries the user for label at every iteration, effectively reducing the number of training data required. This further boosts the prediction performance of a model while guiding through a search space of unlabeled catalyst data.

In this work, we incorporate the above strategies to develop models for predicting oxygen adsorption free energy ($\triangle G_O$) on perovskites. We propose LMTO[9] based geometric features for catalysts that can be extended to materials beyond intermetallics. We use these features along with inexpensive compositional features reported by Li et al.[8b] to develop our prediction models. We then apply the active learning strategies for training our models. A recent study[8b] reports a DFT-based model with root mean square error (RMSE) of 0.6 eV for screening optimal catalyst perovskites from a list of 4000 catalysts for the oxygen evolution reaction. Further studies[13] predict O activity using expensive electronic descriptor with ~0.4 eV mean absolute error (MAE.) Yet another study[14] uses genetic algorithm in symbolic regression to identify empirical relations for accurately predicting activity of perovskites, although, the approach is limited to the knowledge presented in their training set. Herein, we develop more accurate prediction models for a range of perovskites without employing DFT-based input features. Instead, we broaden the implementation of LMTO based geometric features from intermetallics (i.e., materials with active d-bands) to perovskites comprising of s-, p-, d- and f-block elements. For the active learning technique, we implement a committee-based query strategy, that is, maximum disagreement[15] and the expected error reduction strategy.[16] We also outline a workflow that uses time-series forecasting to predict RMSE of unlabeled data based on past iterations. This workflow can be implemented to screen catalysts from search spaces with thousands of possible candidates and can be extended to other reaction chemistries.

Computational details

Dataset

The prediction models are developed for a small sample space of 368 perovskite catalyst materials of the form ABO3 and AA'B2O6 (Pm3 m). The O adsorption energy data for (001)-terminated cubic structures of the perovskites are taken from the study by Li et al..[8b] The prediction models are trained on 75% of the dataset and the rest 25% are used for testing.

Identifying Outliers in the Dataset

The interquartile range (IQR) was used to identify outliers in the data. The IQR of a univariate dataset is computed using Equation (1).

$$IQR = Q_3 - Q_1 \tag{1}$$

where Q_1 represents the 25th percentile of the data and Q_3 represents the 75th percentile of the data. Equation (1) was used to compute the IQR score for every input catalyst feature separately. Based on the IQR score, the corresponding lower (LB) and upper bounds (UB) are defined for each feature, see Equations (2) and (3). If either of the bounds are not satisfied, then the data point is identified as an outlier for that particular catalyst feature.

$$LB = Q_1 - 1.5 \times IQR \tag{2}$$

$$UB = Q_3 + 1.5 \times IQR \tag{3}$$

The data points that were identified as outliers for at least three catalyst feature distributions were defined as the final outliers of the entire multidimensional dataset.Oxygen Adsorption Free Energy Prediction Models and Input Catalyst Features

The scikit-learn library[17] was used to develop the prediction models. Random forest (RFR), extra tree regressions (ETR) and Gaussian process regression (GPR) were trained to predict adsorption free energies on perovskites and their respective RMSEs were evaluated. It was observed that by altering the split of training and test sets, the RMSE/MAE varied within a certain range. For quantitative evaluation, the estimation variance was reduced by repeating the single-shot trial over 100 random test/training splits. The mean of 100 RMSE/MAE estimates was used as the prediction accuracy of the ML model.[18] Further analysis was also performed using 300 and 500 random splits to validate our model evaluation metrics, that is, mean RMSE/MAE. Please refer to Section 1.3 of Supporting Information for more details.

The DFT-based and geometric descriptors-based prediction models were developed and their performances were compared. The DFT-based models employ the complete list of catalyst input features, that is, 18 compositional and 47 DFT-based features, reported by Li et. al..[8b] 45 of the 47 DFT features require DFT calculations for each data-point and the rest two are taken from Materials Project database.[19] Whereas, the geometric based models employ the 15 features adapted from the study by Li et al.[8b] along with two geometric features to define catalysts. Further details on the compositional features can be found in Table S2 of Supporting Information and in Ref [8b]. The LMTO-based geometric features, on the other hand, were evaluated using tight binding linear muffin-tin orbital (TB-LMTO) formulation (Eq. (4)) reported by Harrison and Froyen et al..[9]

$$\mathbf{V}^{\mathbf{BO}}_{\mathbf{ll',m}} = \mathbf{n}_{\mathbf{ll',m}} \frac{\hbar^2}{\mathbf{m_e}} \frac{\left(\mathbf{r}_{\mathbf{B}}^{(2l-1)} \mathbf{r}_{\mathbf{O}}^{(2l'-1)}\right)^{1/2}}{\mathbf{d}_{\mathbf{BO}}^{(l+l'+1)}} \tag{4}$$

where $\mathbf{V}^{\mathbf{BO}}_{\mathbf{ll',m}}$ is the interatomic coupling matrix element between B and O, $\mathbf{l}$ and $\mathbf{l'}$ are the orbital block corresponding to the B and O site elements in the periodic table respectively. It should be noted that $l = 1/2$ for both s and p and $l = 2$ for d-orbital.[20] $\mathbf{n_{ll,m}}$ is a

constant independent of the metal, $\hbar$ is the Planck's constant, $\mathbf{m_e}$ is the mass of an electron, **r** is the spatial extent of the atom's orbital corresponding to **l** and $\mathbf{d_{BO}}$ is the distance between neighboring atoms B and O.

Using Equation (4), we construct the geometric features V_B_O_ sigma and V_B_O_pi, where B corresponds to the site atom and O is the neighboring oxygen atom. The spatial extent $\mathbf{r_B}$ and $\mathbf{r_O}$ for site atom B and O respectively were taken from an online database for elements.[21] Although $\mathbf{d_{BO}}$ requires ab initio computation, herein we estimate $\mathbf{d_{BO}}$ by adding average ionic radii of atom B and O. A total of 17 features were used for training the geometric based RFR and ETR models. Further details on the model hyper parameters can be found in Section 1.1 of the Supporting Information. A sample calculation is presented in Section 1.3 to demonstrate how the geometric features V_B_O_sigma and V_B_O_Pi are computed.

As tree-based regression tends to over-fit with increasing number of input features, a recursive feature elimination using a 4-fold cross validation was performed on the training set to limit this over-fitting.

Active Learning Strategy

The Python modAL library[15] was used to develop the active learning algorithm for our regression models. The query strategies are the expected error reduction and maximum disagreement within a committee. We implemented pool and batch sampling for these strategies. In the pool based sampling, a single data point is added to the training set at every iteration, whereas in the batch sampling, multiple data points are added depending on the batch size (3, 5 and 9).

The algorithm for the expected error reduction strategy is adopted from Douak et al..[16] Herein the active learner labels those points that have the highest expected error at a given iteration. This expected error is obtained from a residual model for each unlabeled data point. The residual model is also a regression model similar to the main prediction model, however, it is trained on the prediction errors of the main model. Further details of the expected error algorithm can be found in Section 2.1 of the Supporting Information.

The algorithm for the committee-based query strategy is shown in Section 2.2 of Supporting Information. Herein a committee is built using two regression models that are initialized on different initial data points. Predictions are made on the unlabeled data points by both of these models and the points having maximum variation in their predicted values are queried. Further details on the query strategies and the corresponding algorithms can be found in Section 2 of the Supporting Information. Python codes implementing these algorithms can be accessed via the link mentioned in the Supporting Information.

The active learning algorithm is allowed to iteratively add labelled data points to the training set, until the training set is 75% of the total available dataset. We understand that this termination criterion is difficult to be followed when a dataset contains thousands of data points; for example the test data provided by Li et al.[8b] consists of 4,000 unlabeled perovskites data points. Therefore, to identify the termination criteria for huge datasets, we employ the forecasting method Autoregressive Integrated Moving Average (ARIMA)[22] with batch sampling to predict the batch RMSE at the end of each iteration. The batch RMSE helps us estimate the actual prediction accuracy of the model and the forecast identifies and predict trends in the batch RMSE for subsequent iterations. The batch RMSE e_b is given by Equation (5)

$$e_b = \sum_{i}^{n} \left(y_i - \widehat{f}(x_i)\right)^2 \tag{5}$$

where, n is the batch size, $\widehat{f}(x_i)$ is the model prediction for x_i and y_i is the label.

Batch RMSEs of the first 20 iterations are used to train the time-series forecasting model ARIMA. Further details of the ARIMA model parameters are given in Section 3.3 of the Supporting Information.

2. Results and Discussion

The model accuracies for predicting O adsorption free energies ($\triangle G_O$) on the B-site (100)-terminated perovskite structures are reported here. We first report the prediction accuracies as obtained from random splitting of test/train datasets, followed by the active learning based sampling of the training data. Lastly, we forecast the batch RMSE based on the initial 20 iterations and compare the prediction results with the actual RMSE to justify our workflow.

2.1. O Adsorption Free Energy Prediction Models

Table 1 shows the mean test RMSEs of the RFR, ETR and GPR regression models trained on 100 randomly split training datasets. Figure 1 shows parity plots for RFR and ETR methods. Further details of the prediction models are given in Section 2.3 of Supporting Information. It was found that the mean RMSE/MAE remains unchanged as we increase the number of random test/train splits from 100 to 500. Please refer to Section 1.4 of Supporting Information for more details regarding test/train splits and Section 1.5 for baseline comparison, that is, models trained on geometric features except the two LMTO features.

Table 1. Mean RMSEs/MAEs of random forest regression (RFR), extra tree regression ETR and Gaussian process regression models trained on DFT based features[8b] and geometric features for prediction of DFT-calculated adsorption free energies of O ($\triangle G_O$) as obtained from 100 random training/test splits.

		Training error [eV] (RMSE/MAE)		Test error [eV] (RMSE/MAE)			
Model	Feature	Mean	Std. dev.	Mean	Std. dev.	Min.	Max.
RFR	Geometric	0.32/0.22	0.02/0.01	0.78/0.54	0.11/0.06	0.57/0.41	1.01/0.70
	DFT-based	0.30/0.20	0.02/0.01	0.75/0.52	0.11/0.06	0.51/0.40	1.04/0.65
ETR	Geometric	0.08/0.05	0.01/0.00	0.65/0.44	0.09/0.05	0.50/0.33	0.91/0.57
	DFT-based	0.04/0.02	0.00/0.00	0.58/0.40	0.09/0.05	0.40/0.30	0.89/0.55
GPR	Geometric	0.52/0.36	0.05/0.03	0.95/0.67	0.11/0.06	0.73/0.55	1.30/0.87
	DFT-based	0.25/0.18	0.07/0.04	0.63/0.46	0.07/0.05	0.47/0.34	0.81/0.57

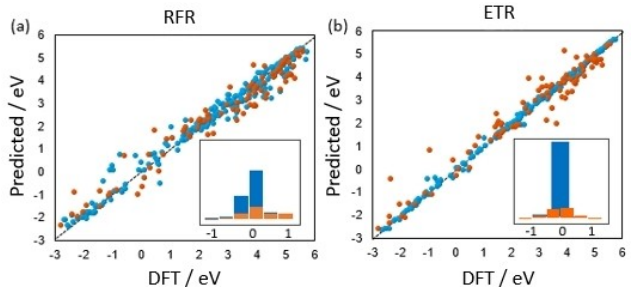

Figure 1. Parity plots of RFR (a) and ETR (b) models for prediction of DFT-calculated adsorption free energies of O ($\triangle G_O$) using geometric input catalyst features. Color code: blue corresponds to the training set (75%), red corresponds to the test set (25%).

Unlike the RFR model, both ETR and GPR models have significantly better test prediction errors with the DFT-based features rather than with the geometric features. The ETR model has the least prediction errors among the regression models for either of the input features. However, after comparing the training and test errors in Table 1, we find that ETR model is over-fitting on the training data, the extent of which increases with the number of input features, that is, from 17 geometric features to 66 DFT based. Even for the geometric input features the ETR model is highly over fitted. By observing the parity plots of the RFR and ETR models in Figure 1, we see that the ETR's prediction on the training data is very accurate whilst its test data predictions are not. Whereas in the case of the RFR model, we see that both the test and the train predictions are almost equally spread along the line $y = x$, further suggesting overfitting of the ETR model.

The DFT-based GPR model significantly out performs the corresponding geometric feature model predictions. However, DFT based features are computationally expensive, therefore this study only focuses on models that provide reasonable prediction errors with geometric features, that is, ETR and RFR. Given that overfitting issue is common with tree based models that are trained on limited data points with high numbers of input features,[18] in the later sections we will demonstrate how active learning strategy can improve the prediction errors while limiting the overfitting problem.

2.2. Active Learning Based Prediction Models

The mean RMSEs/MAEs for 100 randomly initialized active learning based RFR and ETR regression models are reported in this section. Table 2 shows the prediction performance of different query strategies for the pool-based sampling. It should be noted that the catalyst features of the unlabeled catalyst data points are accessed by the query strategy of the active learning algorithm. However, the test set is always defined as an unseen distribution. Thus, prediction errors of the unlabeled set are labelled as validation set RMSE/MAE instead of test RMSE/MAE.

From Table 2 we observe that the active learning strategy significantly improves the mean RMSE/MAE of the regression models. It is further found that the committee-based strategy performs better than the expected error strategy for both regression models, especially for the ETR model, where a mean validation RMSE/MAE of 0.40/0.25 eV is achieved and the std. dev. is within 0.09/0.04. These prediction errors are significantly better than the currently reported prediction errors[8b] of 0.8 eV and 0.5 eV obtained from the corresponding adaptive learning models employing 18 compositional features and 18 compositional plus 47 DFT-based features respectively. Hence, we perform a further batch-based sampling for the committee based query strategy. Table 3 shows the mean RMSE/MAE for the batch sampling of training data with different batch sizes.

After comparing the mean test and train errors in Table 3, we observe that the overfitting of ETR resolves as we increase the batch size. The overfitting is further reduced by employing recursive feature elimination (RFE) based on 4-fold cross validation of the training set at every iteration. Using the RFE technique we also identify that the LMTO-based features are among the most important features (based on feature ranking). Please refer to Section 2.1 for more details. Even though the mean RMSE/MAE increases with the batch size and subsequent feature elimination, the mean validation RMSEs/MAEs of the RFR and EFR remain within 0.6/0.5 eV and 0.5/0.4 eV respectively.

The maximum RMSEs/MAEs are around 0.7/0.5 eV for the ETR model. These high prediction errors occur when the models in the committee turn out to be similar due to bad initialization. We can increase the probability of the models turning out to be dissimilar by initializing them on the outliers, thus further reducing the mean validation RMSE/MAE.

The above initialization strategy was tested for committee based query on both ETR and RFR models with a batch size of

Table 2. Mean RMSEs/MAEs of three different active learning query strategies employed via pool-based sampling on the RFR and ETR regression models using geometric features for prediction of the DFT-calculated adsorption free energies of O ($\triangle G_O$). The errors are averaged over of 100 random initializations.

		Training error [eV] (RMSE/MAE)		Validation error [eV] (RMSE/MAE)			
	Query strategy	Mean	Std. dev.	Mean	Std. dev.	Min.	Max.
RFR	Expected error (test)	0.32/0.22	0.01/0.01	0.57/0.41	0.09/0.05	0.43/0.31	0.72/0.50
	Expected error (train)	0.31/0.21	0.01/0.01	0.53/0.38	0.08/0.05	0.37/0.29	0.69/0.48
	Max. dis-agreement within committee	0.33/0.23	0.01/0.01	0.47/0.35	0.06/0.03	0.37/0.26	0.68/0.46
ETR	Expected error (test)	0.09/0.05	0.01/0.00	0.52/0.35	0.11/0.05	0.37/0.25	0.86/0.49
	Expected error (train)	0.09/0.05	0.01/0.00	0.49/0.34	0.09/0.04	0.35/0.25	0.77/0.44
	Max. dis-agreement within committee	0.11/0.06	0.01/0.00	0.40/0.25	0.09/0.04	0.26/0.19	0.62/0.36

Table 3. Mean prediction RMSE/MAE over 100 random initializations of RFR and ETR models trained via batch sampling of maximum disagreement strategy. The batch sizes 3, 5 and 9 are investigated followed by models trained with recursive feature elimination (RFE) at each active learning iteration for batch size 9.

		Training [eV] (RMSE/MAE)		Validation error [eV] (RMSE/MAE)			
	Batch Size	Mean	Std. dev	Mean	Std. dev.	Min.	Max.
RFR	3	0.33/0.23	0.02/0.01	0.51/0.37	0.07/0.04	0.35/0.29	0.76/0.51
	5	0.34/0.23	0.02/0.01	0.54/0.39	0.07/0.04	0.40/0.32	0.76/0.51
	9	0.35/0.24	0.02/0.01	0.54/0.40	0.07/0.04	0.37/0.29	0.70/0.52
	9 (RFE)	0.38/0.26	0.02/0.01	0.57/0.40	0.09/0.05	0.41/0.29	0.89/0.53
ETR	3	0.12/0.06	0.02/0.00	0.43/0.29	0.09/0.05	0.28/0.20	0.65/0.40
	5	0.13/0.06	0.02/0.00	0.44/0.30	0.08/0.04	0.31/0.22	0.63/0.42
	9	0.15/0.07	0.03/0.01	0.48/0.32	0.07/0.04	0.33/0.25	0.64/0.42
	9 (RFE)	0.23/0.14	0.04/0.03	0.47/0.32	0.08/0.04	0.31/0.23	0.66/0.41

9. Outlier initialization further reduced the mean prediction RMSE/MAE to 0.45/0.31 eV and 0.54/0.40 eV for ETR and RFR models respectively. Please refer to Section 3.2 of Supporting Information for more details.

We also computed the performance of DFT-based models when trained with the committee-based query for a batch size of 9 and compared it with geometric features based models. We found that geometric feature-based models provide slightly better prediction accuracy than that obtained from DFT-based model. Please refer to Table S6 of Supporting Information for more details.

2.3. Batch RMSE Forecasting

The idea behind computing batch RMSE is to get an estimate on the actual prediction accuracy of the model, that is, actual RMSE of the unlabelled data points without having to label them. Figure 2 shows this relation between batch RMSE and actual RMSE of the unlabelled data as a function of the number of iterations.

From Figure 2 we observe that the batch RMSE fluctuates around the actual RMSE. These fluctuations are quite high for small batch sizes, since only a few data points are considered while evaluating batch RMSE. However, as the batch size increases, the batch RMSE curve begins to flatten out giving a better estimate of the actual RMSE, that is, the confidence interval for predicting the actual RMSE narrows out.

For problems involving bigger search spaces, like the test set of 4,000 perovskites provided by Li et al,[8b] a bigger batch size can be defined to further narrow the confidence interval for predicting the actual RMSE. This again allows us to define a termination criteria for an iterative active learning algorithm based on achieving an acceptable batch RMSE within the confidence interval, that is, when fluctuations in batch RMSE start to converge.

Batch RMSEs can also be used to train forecasting models, like ARIMA, that can further reduce the number of expensive label evaluations. Figure 3 shows this forecasting technique to predict batch RMSE for a batch size of 9, given the available dataset. The first 20 iterations were used for computing the starting parameters of the ARIMA prediction model. Further details of the model can be found in Section 3.4 of Supporting Information.

From Figure 3 we observe that a classical time series forecasting method like ARIMA can provide a near accurate prediction of batch RMSE (batch size 9) for up to 10 iterations, that is, 90 data points without having to label them.

For search spaces with thousands of unlabeled data points, a batch size as big as 20–30 data points can be defined to eliminate the fluctuations, thus reducing the training size requirement from thousands to only a few hundreds of labeled data points.

The forecast method can also help pre-allocate computational resources beforehand by predicting the number of iterations before termination, that is achieving required pre-

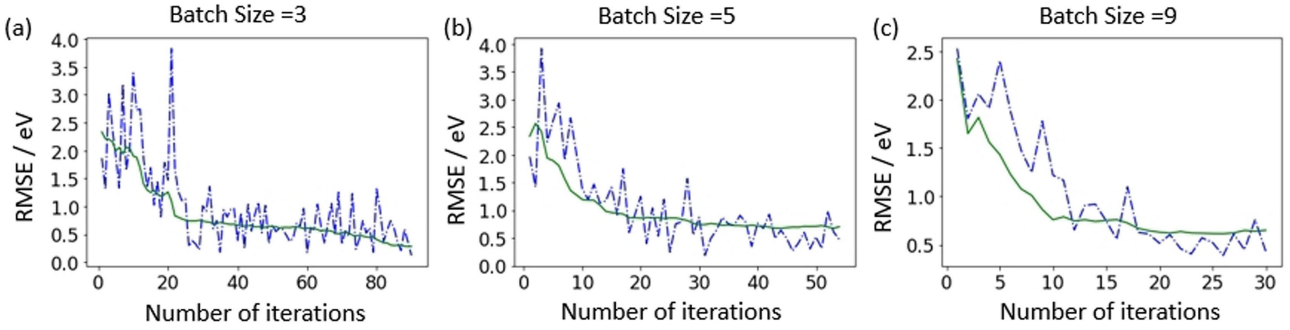

Figure 2. Batch RMSE (blue dotted line) and actual RMSE (green) vs. the number of iterations for batch sizes 3 (a), 5 (b) and 10 (c). The number of iterations are such that the validation set size remains 25% of the total dataset for every batch size.

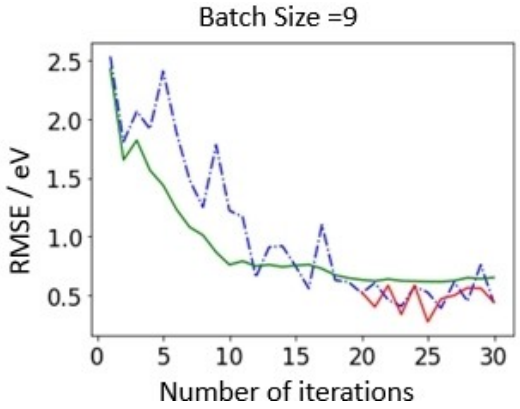

Figure 3. Batch RMSE forecasting using ARIMA trained on first 20 iterations. Blue dotted line corresponds to the batch RMSE at every iteration, red line is the predicted batch RMSE for subsequent iterations and green line is the actual RMSE for predicting unlabelled data points at every iteration.

diction accuracy or error convergence. However, termination based on ARIMA forecasting should be avoided. This is because ARIMA forecast works well for short term forecasting, that is the accuracy of the forecast decreases as the number of subsequent iteration predictions increases. There are forecasting tools that outperforms ARIMA in long term forecasting like the 'Long short-term memory' (LTSM)[23] which is based on a recurrent neural network architecture. However, these tools have a more complex build than ARIMA. Thus, identifying the relevant of forecasting tool for a given dataset is necessary but beyond the scope of this paper.

3. Conclusions

We present a workflow for developing prediction models using active learning and forecasting techniques, which is designed to reduce the number of expensive data labelling evaluations. These prediction models are built using catalysts features that are inexpensive compositional and geometric features based on linear muffin-tin orbital theory. The features can be extended to search spaces consisting of catalysts beyond d-block metals. Herein, we demonstrate their implementation for a range of perovskites catalysts comprising of s-, p-, d- and f-block elements.

The workflow is based on the algorithms for expected error and committee-based query strategies that can be extended to any problems that require expensive data labelling. The pool-based sampling of ETR models have the lowest prediction errors, that is mean validation RMSE/MAE for the committee based query over 100 random initialization is 0.40/0.25 eV. However, the batch based sampling followed by recursive feature elimination reduces overfitting in the extra tree regression model while providing a reasonable prediction accuracy without employing DFT based input features, that is mean validation RMSE/MAE of 0.47/0.32 eV for a batch size of 9, over 100 random initializations. The errors suggest that the model performance is better than the current DFT based models in the literature, whereby a models with prediction RMSE <0.6 eV and MAE <0.4 eV are employed for screening of perovskites for the oxygen evolution reaction.

We further demonstrate that batch RMSE can be used to estimate the actual prediction accuracy of the model, that is actual prediction RMSE of unlabeled data points without having to label them. The confidence interval of this estimate can be narrowed by increasing the batch size because fluctuation in the batch RMSE decreases with bigger batch sizes. We then employ the forecasting method to predict batch RMSEs of subsequent iterations, thus reducing the data labelling requirement for up to 90 data points for a training set of 276 data points. This methodology helps to effectively allocate computational resources for training prediction models, given the limited availability of labeled data points.

4. Supporting Information Summary

The Supporting Information provides the following details: 1. Regression model parameters 2. Active learning strategies (details including the algorithm) 3. Batch Sampling, feature ranking, outlier initialization, for comparison with DFT-based models and forecasting (ARIMA model parameters).

Python codes for implementing the active learning algorithms can be accessed via the GitHub repository:

https://github.com/shambhawipandey29/Active-learning-implementations.

Acknowledgements

S.S. acknowledges the research scholarship funding from Science and Engineering Research Board, India and Cambridge Trust. This work was in part funded by UKRI Centre for Doctoral Training "Automated Chemical Synthesis Enabled by Digital Molecular Technologies" EP/S024220/1.

Conflict of Interest

The authors declare no conflict of interest.

Keywords: active learning · adsorption energy prediction · electrochemistry · heterogeneous catalysis · perovskites

[1] a) O. Mohan, Shambhawi, A. A. Lapkin, S. H. Mushrif, *Catal. Sci. Technol.* **2020**, *10*, 6628–6643; b) T. P. de Carvalho, R. C. Catapan, A. A. M. Oliveira, D. G. Vlachos, *Ind. Eng. Chem. Res.* **2018**, *57*, 10269–10280; c) O. Mohan, S. Shambhawi, R. Xu, A. A. Lapkin, S. H. Mushrif, *ChemCatChem* **2021**, *13*, 2420–2433.
[2] C. T. Campbell, *ACS Catal.* **2017**, *7*, 2770–2779.
[3] T. Bligaard, J. K. Nørskov, S. Dahl, J. Matthiesen, C. H. Christensen, J. Sehested, *J. Catal.* **2004**, *224*, 206–217.
[4] a) M. M. Montemore, J. W. Medlin, *Catal. Sci. Technol.* **2014**, *4*, 3748–3761; b) J. Greeley, *Annu Rev Chem Biomol Eng* **2016**, *7*, 605–635.

[5] a) Z. Li, X. Ma, H. Xin, *Catal. Today* **2017**, *280*, 232–238; b) Z. Li, S. Wang, W. S. Chin, L. E. Achenie, H. Xin, *J. Mater. Chem. A* **2017**, *5*, 24131–24138; c) J. Noh, S. Back, J. Kim, Y. Jung, *Chem. Sci.* **2018**, *9*, 5152–5159.
[6] a) S. Back, J. Yoon, N. Tian, W. Zhong, K. Tran, Z. W. Ulissi, *J. Phys. Chem. Lett.* **2019**, *10*, 4401–4408; b) T. Xie, J. C. Grossman, *Phys. Rev. Lett.* **2018**, *120*, 145301.
[7] F. Abild-Pedersen, J. Greeley, F. Studt, J. Rossmeisl, T. R. Munter, P. G. Moses, E. Skúlason, T. Bligaard, J. K. Nørskov, *Phys. Rev. Lett.* **2007**, *99*, 016105.
[8] a) M. Andersen, S. V. Levchenko, M. Scheffler, K. Reuter, *ACS Catal.* **2019**, *9*, 2752–2759; b) Z. Li, L. E. K. Achenie, H. Xin, *ACS Catal.* **2020**, *10*, 4377–4384.
[9] W. A. Harrison, S. Froyen, *Phys. Rev. B: Condens. Matter Mater. Phys.* **1980**, *21*, 3214–3221.
[10] L. Chanussot, A. Das, S. Goyal, T. Lavril, M. Shuaibi, M. Riviere, K. Tran, J. Heras-Domingo, C. Ho, W. Hu, A. Palizhati, A. Sriram, B. Wood, J. Yoon, D. Parikh, C. L. Zitnick, Z. Ulissi, arXiv preprint arXiv:2010.09435, **2020**.
[11] T. Toyao, Z. Maeno, S. Takakusagi, T. Kamachi, I. Takigawa, K.-i. Shimizu, *ACS Catal.* **2020**, *10*, 2260–2297.
[12] K. Tran, Z. W. Ulissi, *Nat. Catal.* **2018**, *1*, 696–703.
[13] a) C. F. Dickens, J. H. Montoya, A. R. Kulkarni, M. Bajdich, J. K. Nørskov, *Surf. Sci.* **2019**, *681*, 122–129; b) W. T. Hong, R. E. Welsch, Y. Shao-Horn, *J. Phys. Chem. C* **2016**, *120*, 78–86.
[14] B. Weng, Z. Song, R. Zhu, Q. Yan, Q. Sun, C. G. Grice, Y. Yan, W.-J. Yin, *Nat. Commun.* **2020**, *11*, 3513.
[15] T. Danka, P. Horvath modAL: A modular active learning framework for Python, can be found under https://github.com/modAL-python/modAL.
[16] F. Douak, F. Melgani, N. Benoudjit, *Appl. Energy* **2013**, *103*, 328–340.
[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
[18] T. Toyao, K. Suzuki, S. Kikuchi, S. Takakusagi, K.-I. Shimizu, I. Takigawa, *J. Phys. Chem. C* **2018**, *122*, 8315–8326.
[19] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, *APL Mater.* **2013**, *1*, 011002.
[20] O. K. Andersen, O. Jepsen, *Phys. Rev. Lett.* **1984**, *53*, 2571–2574.
[21] M. Winter, *WebElements*, can be found under https://www.webelements.com.
[22] R. J. Hyndman, G. Athanasopoulos, *Forecasting: Principles and Practice*, 2nd edition, OTexts: Melbourne, Australia, **2014**.
[23] A. Azari, P. Papapetrou, S. Denic, G. Peters, in *Discovery Science* (Eds.: P. Kralj Novak, T. Šmuc, S. Džeroski), Springer International Publishing, Cham, **2019**, pp. 129–144.

Manuscript received: April 28, 2021

FULL PAPERS

The study reports geometric features that can be generalized to catalyst compositions beyond d-band metals. By employing these features into an active learning workflow, required prediction accuracies are achieved without expensive an-initio features. The workflow itself enables effective data labelling, but feeding it into a forecasting model further limits the number of expensive data labelling.